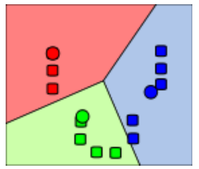
From (chemical) data to information

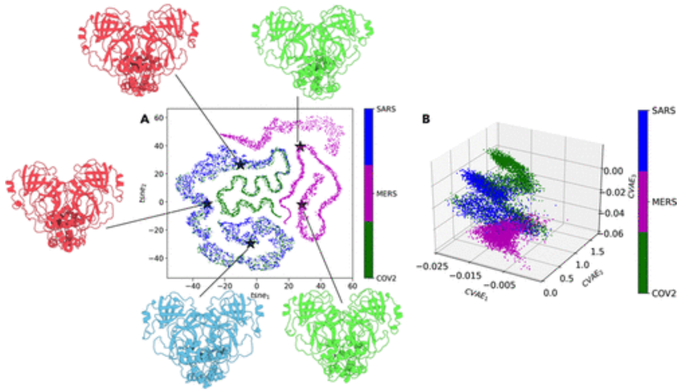
Introduction to Machine learning for Chemistry Workshops 1

Dr Antonia Mey

antonia.mey@ed.ac.uk

Room 214 JBB





Plan for the workshops

Workshop 1 (MH -)

- What are molecular simulations?
- Dimensionality reduction examples for
  - PCA, tICA, t-SNE
- Applied to Molecular dynamics

Workshop 2 (JCMB

- Clustering data
- Classification of data

Workshop 3

- Calssification with Multilayer perceptrons
- Introduction to PyTorch and Multilayer perceptrons

Small Project: 10th March 5pm

- Apply everything you have learned to chemistry example
- Submit a project report in form of a Jupyter notebook.
  - Marking criteria:
    - Does the code run
    - Is the code documented
    - Do you use Markdown to make sure the report flows
    - Are plots labelled correctly and clear do read
    - Include a conclusion and discussion

Plan for today

Workshop content can be found here:

https://github.com/Edinburgh-Chemistry-Teaching/ML-for-Chemistry

```
$: git pull origin main
```

Setting up your environment:

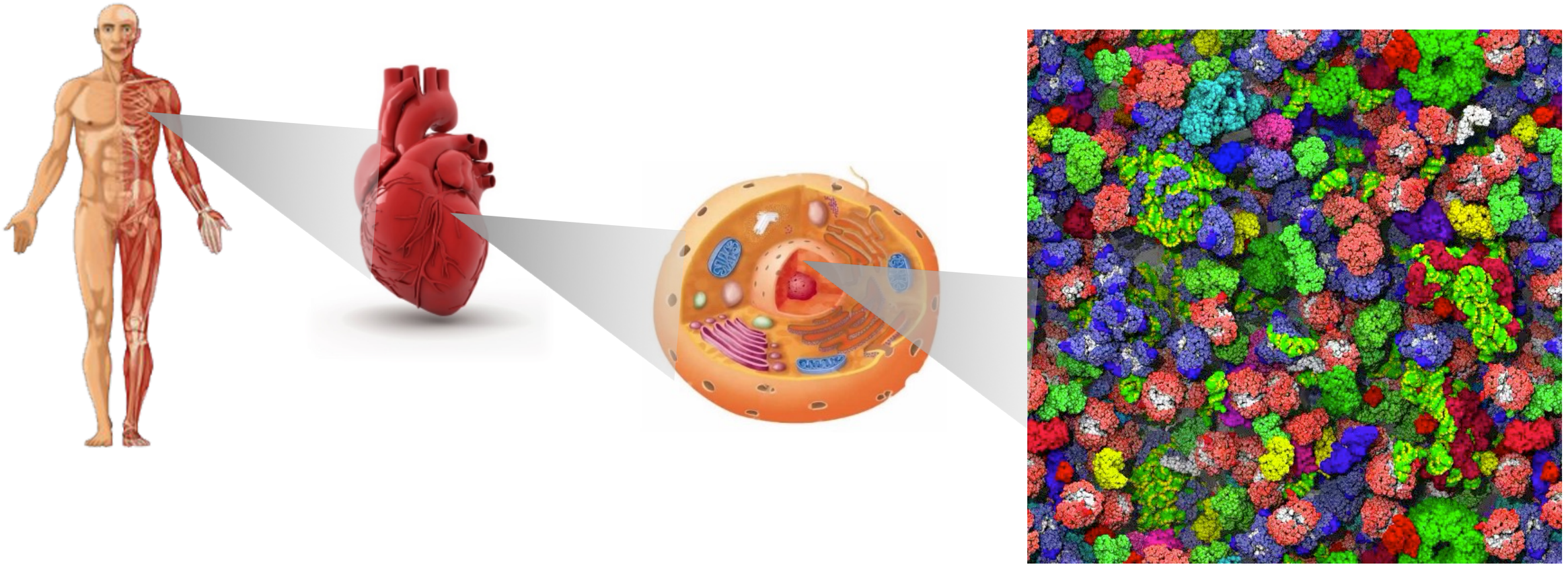
```
$: conda create --file=environment.yml
```

- Brief introduction to molecular simulations
- Run the workshop notebooks

Additional reading:

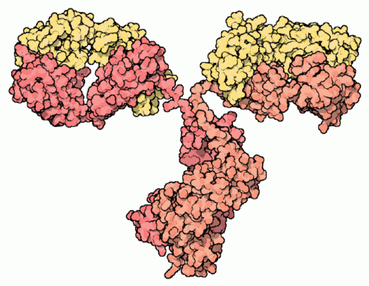
https://dmol.pub/ml/introduction.html

Structure determines function

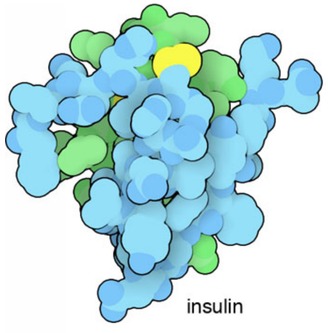


Proteins are molecular machines managing life

Antibody

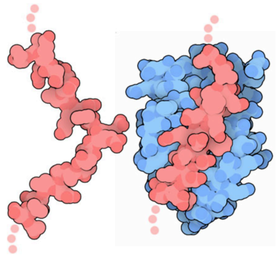


disease



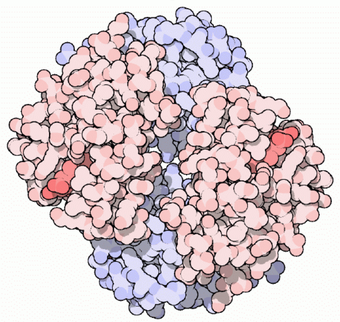
hormones

NOXA



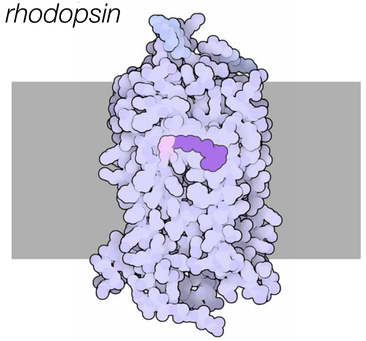
Cell regulation

hemoglobin



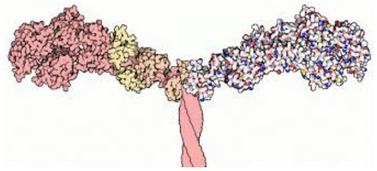
transport

rhodopsin



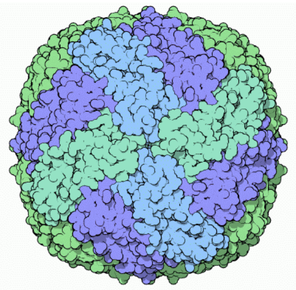
receptors

myosin



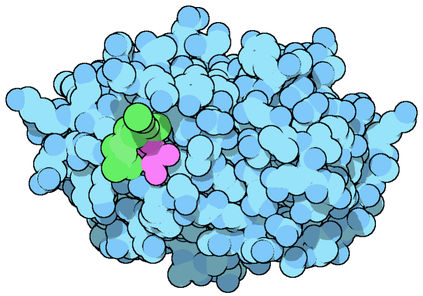
energy

ferritin



storage

trypsin

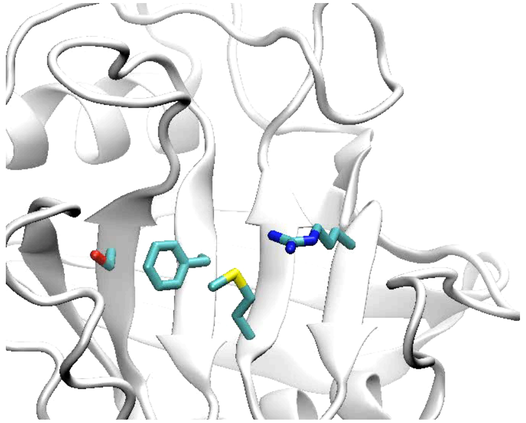


enzymes

https://pdb101.rcsb.org/

What are molecular simulations?

Cyclophilin



Catalysis

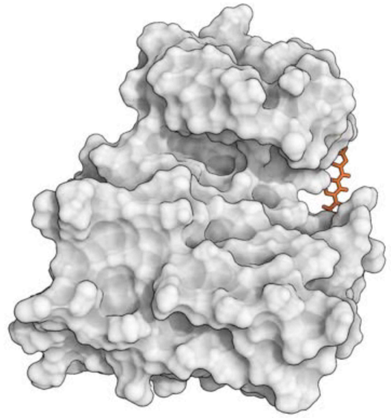
Wapeesittipan, **Mey**, et al., *Comms. Chem.* **2**, 41 (2019)

45±1%
1±1%

55±1%
99±1%

Populations/ rates

Tyrosine kinase —dasatanib



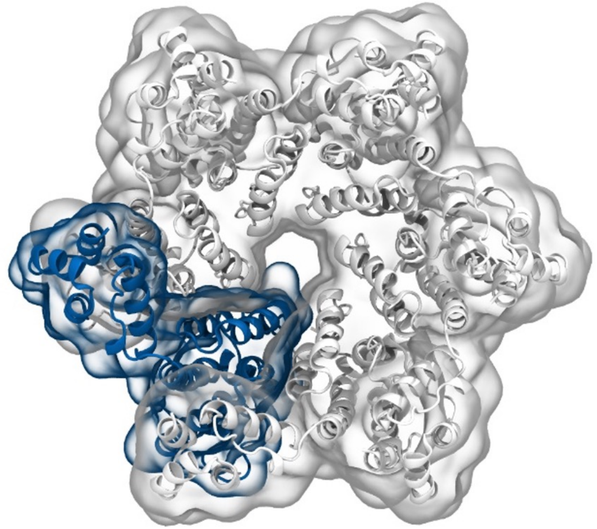
Binding affinities

Shan Y, et al. *JACS* **133,** 9181 (2011)

ΔG k_{on}

Free energies / rates

HIV Capsomer



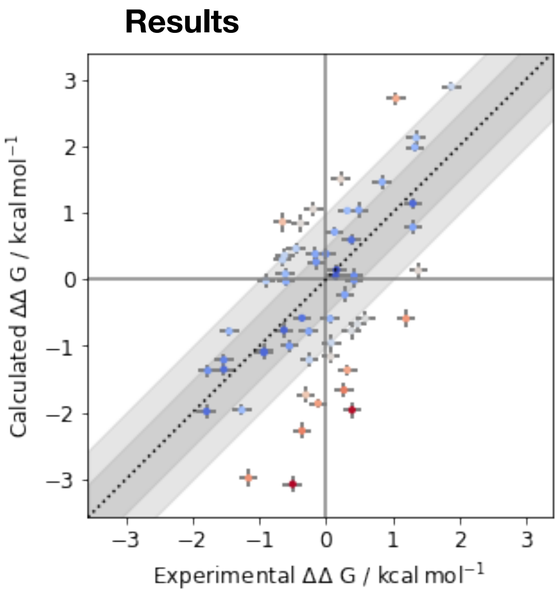
Assembly

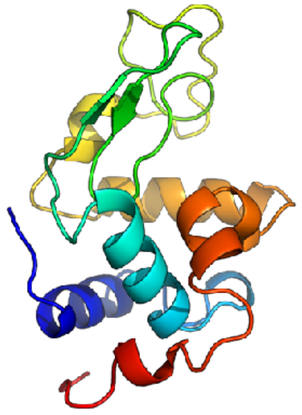
Degiacomi, *Structure* **27**, 1034 (2019)

ΔG

Free energies

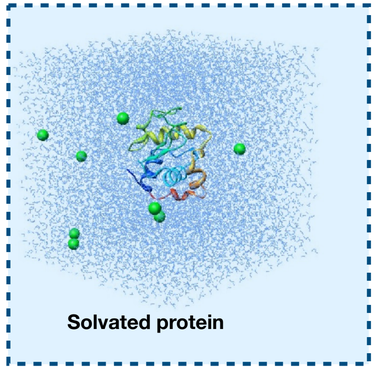
Molecular dynamics simulations can provide quantitative observables



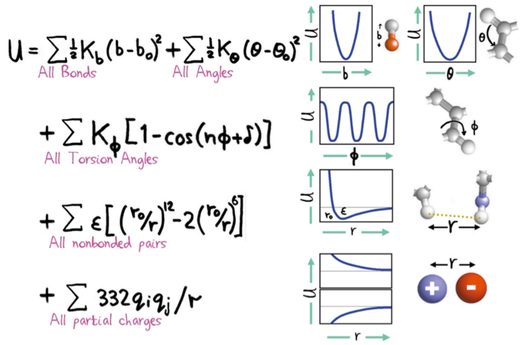


Disulphide bridges
Protonation
Crystal Waters
Cofactors
Alternative side chains

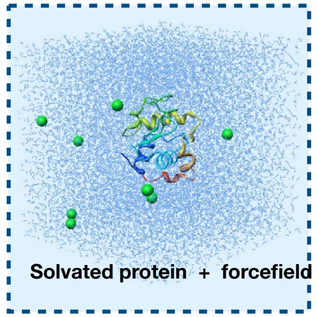
Solvate in water and add ions



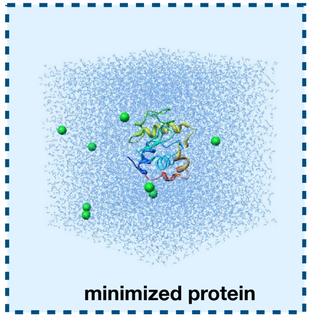
$$\vec{F}_i = -\frac{dU}{d\vec{x}_i}$$



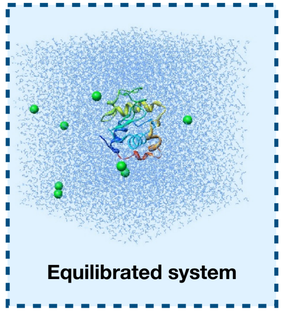
Apply forcefield description

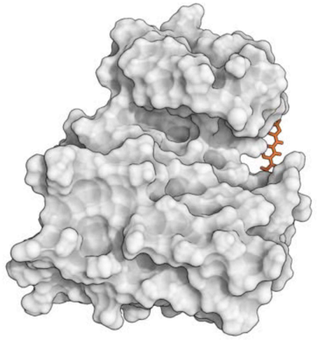


Minimize the positions



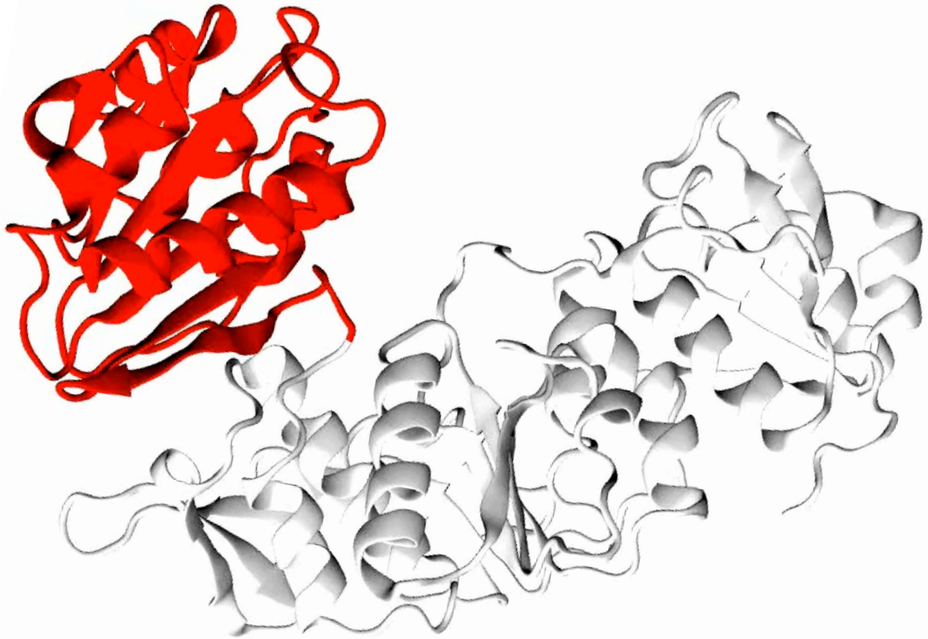
NVT/NPT equilibration

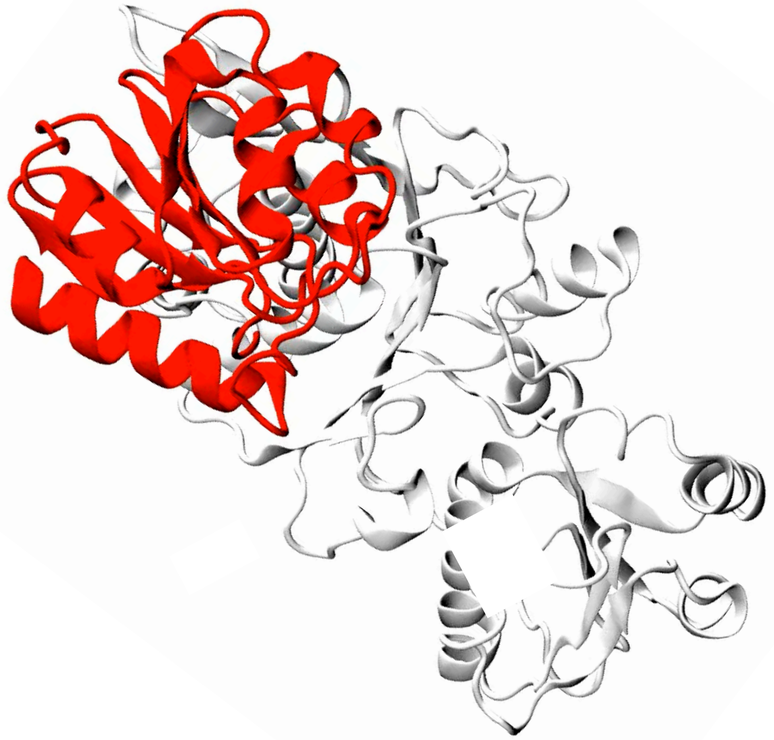




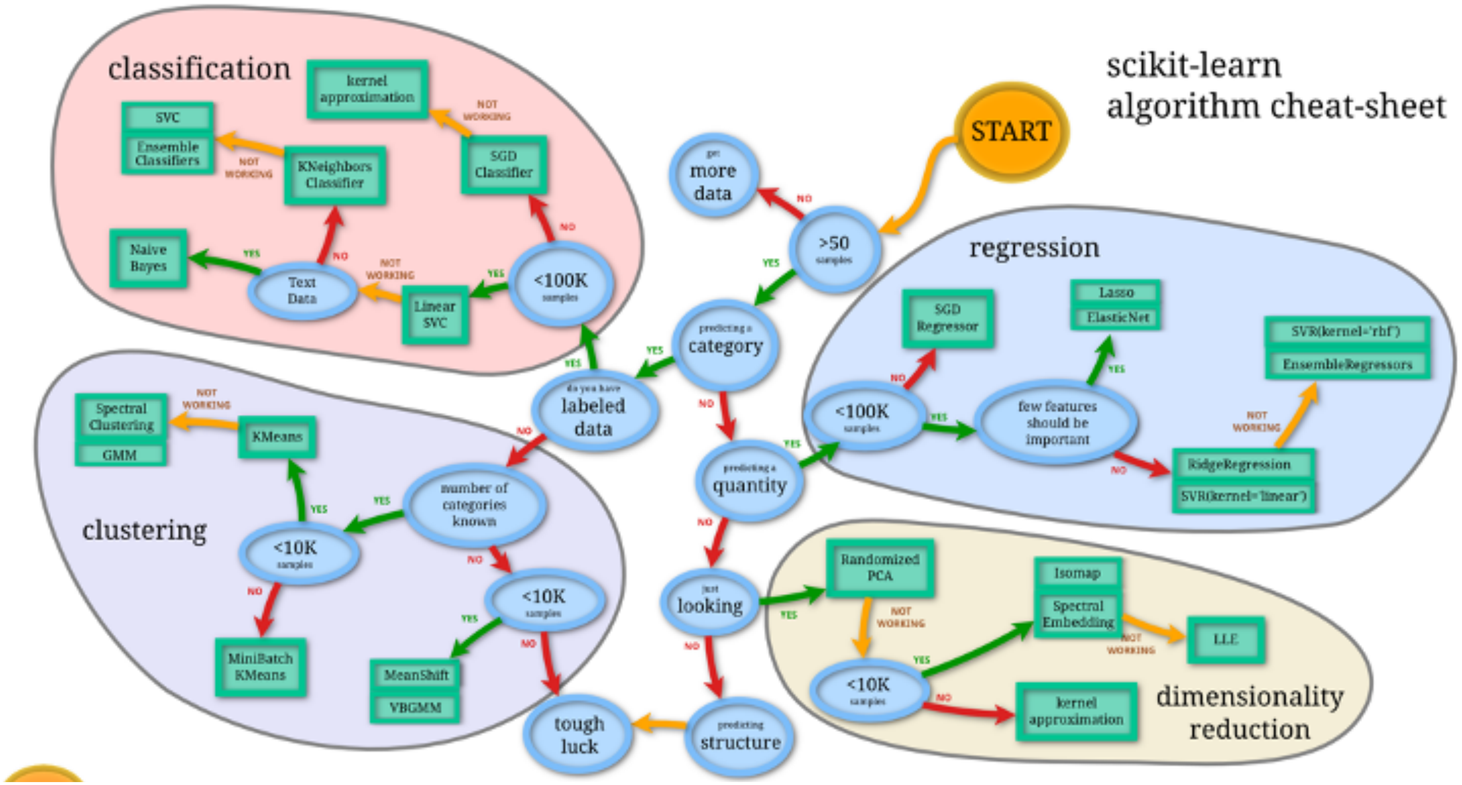
Production simulation

Defining coordinates that describe movement in proteins

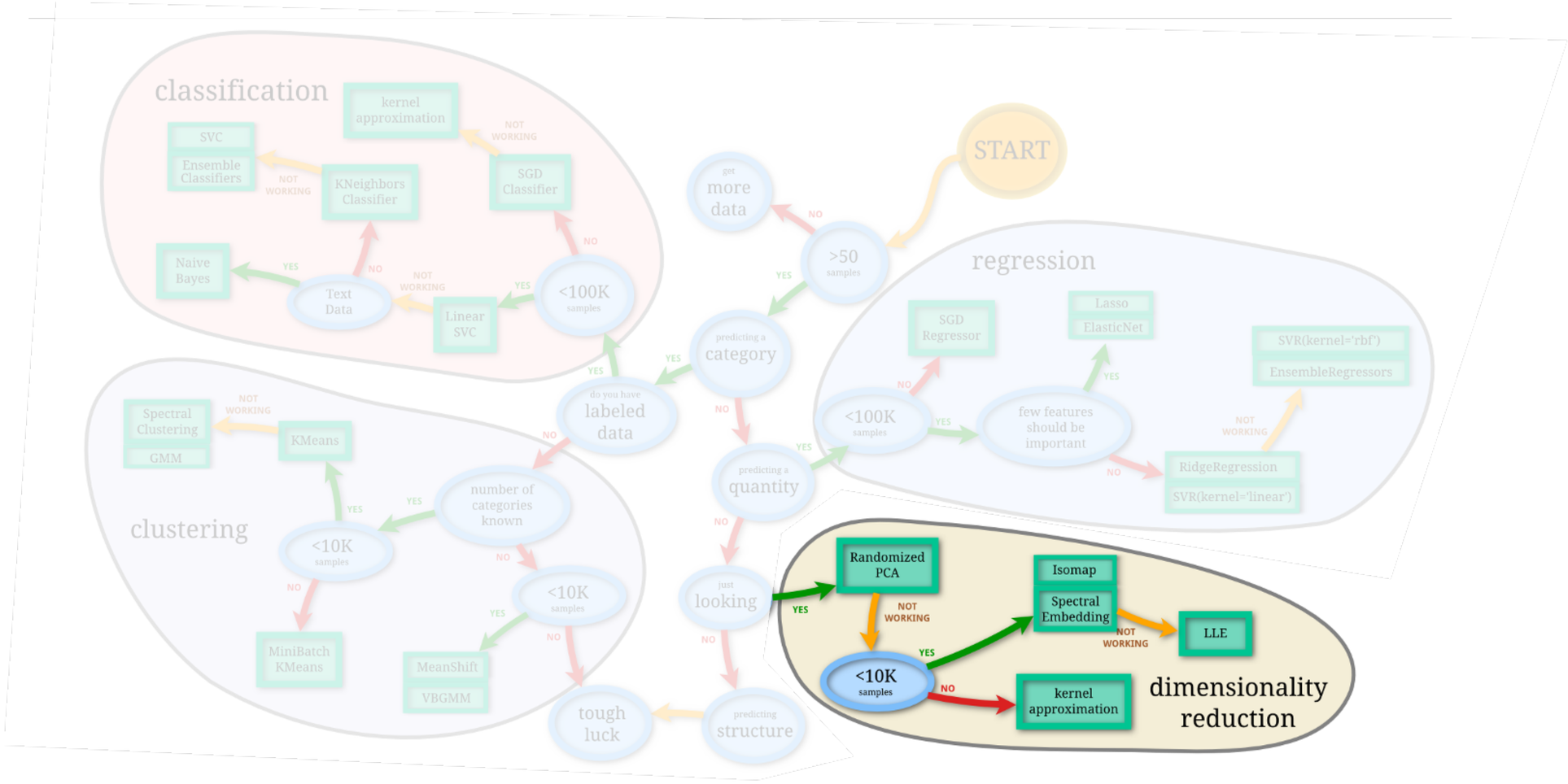




Recall from the lectures: The Data Mining World

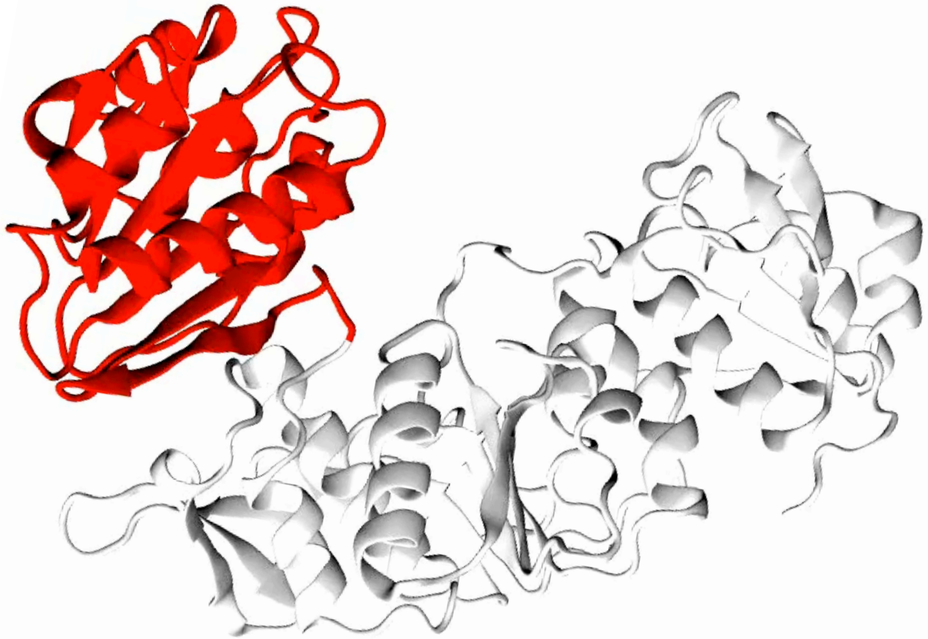


The Data Mining World — Dimensionality Reduction

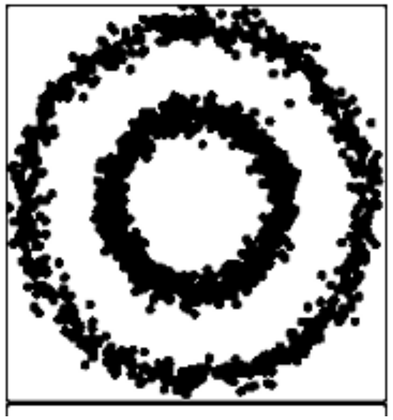


From **scikit-learn.org**

What are the ‘dominant’ dimensions?



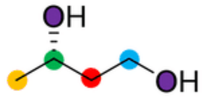
≠



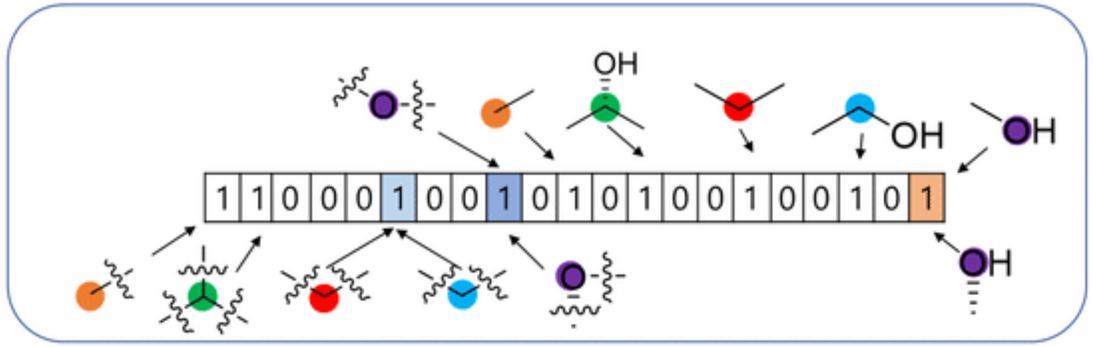
Dimensionality reduction

Feature vectors

```
['ATOM:ACE 1 CH3 1 x',
 'ATOM:ACE 1 CH3 1 y',
 'ATOM:ACE 1 CH3 1 z',
 'ATOM:ACE 1 C 4 x',
 'ATOM:ACE 1 C 4 y',
 'ATOM:ACE 1 C 4 z',
 'ATOM:ACE 1 O 5 x',
 'ATOM:ACE 1 O 5 y',
 'ATOM:ACE 1 O 5 z',
 'ATOM:ALA 2 N 6 x',
 'ATOM:ALA 2 N 6 y',
 'ATOM:ALA 2 N 6 z',
 'ATOM:ALA 2 CA 8 x',
 'ATOM:ALA 2 CA 8 y',
 'ATOM:ALA 2 CA 8 z',
 'ATOM:ALA 2 CB 10 x',
 'ATOM:ALA 2 CB 10 y',
 'ATOM:ALA 2 CB 10 z',
 'ATOM:ALA 2 C 14 x',
 'ATOM:ALA 2 C 14 y',
 'ATOM:ALA 2 C 14 z',
 'ATOM:ALA 2 O 15 x',
 'ATOM:ALA 2 O 15 y',
 'ATOM:ALA 2 O 15 z',
 'ATOM:NME 3 N 16 x',
 'ATOM:NME 3 N 16 y',
 'ATOM:NME 3 N 16 z',
 'ATOM:NME 3 C 18 x',
 'ATOM:NME 3 C 18 y',
 'ATOM:NME 3 C 18 z']
```



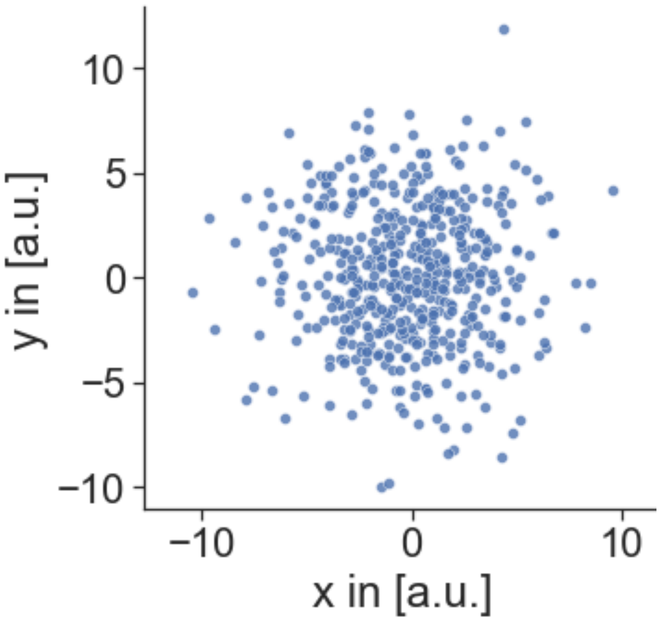
Molecular structure

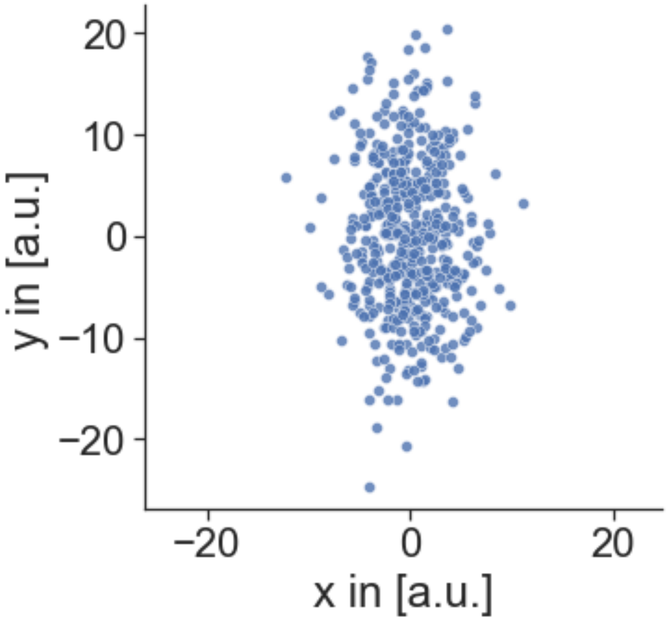


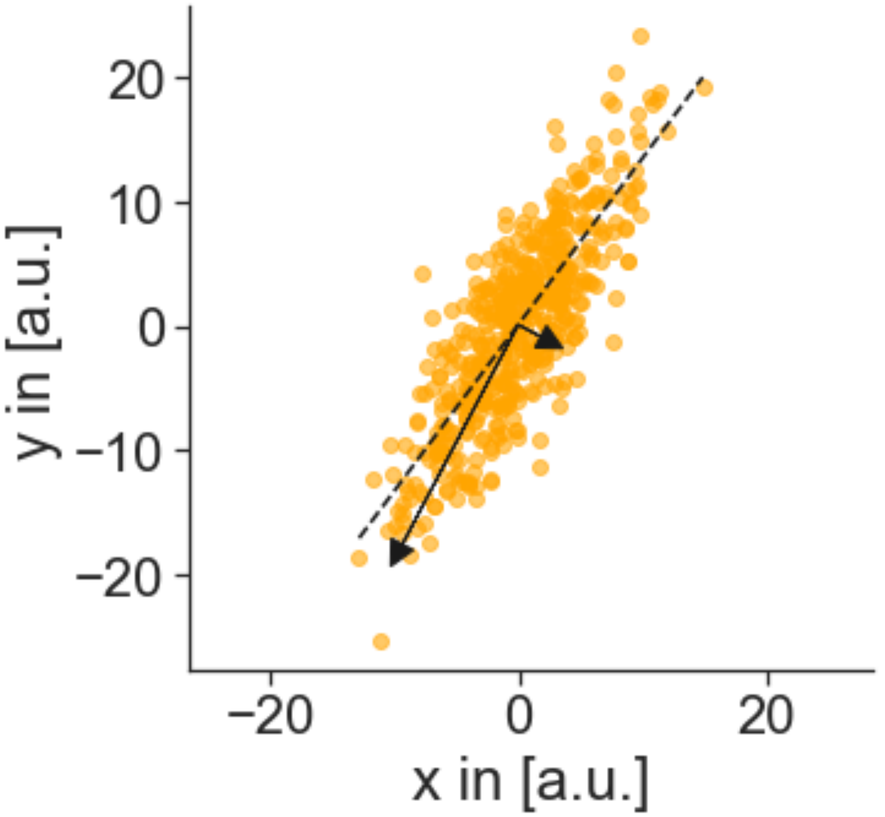
Original Morgan fingerprints

- Clustering on high dimensional data is difficult because often **not enough data is available.**
- How can you identify the **most important features before e.g. clustering?**

Principal component analysis (PCA)







- PCA is an **orthogonal linear transformation** that **maximises the variance** across the first component
- A linear regression fit **minimises the error** with regard to all data points.
- PCA can be used as a tool for **dimensionality reduction**

Time-lagged independent component analysis

- tICA is a **linear transform** similar to PCA
- The transform is chosen such that amongst all linear transforms, tICA **maximizes the autocorrelation** of transformed coordinates.

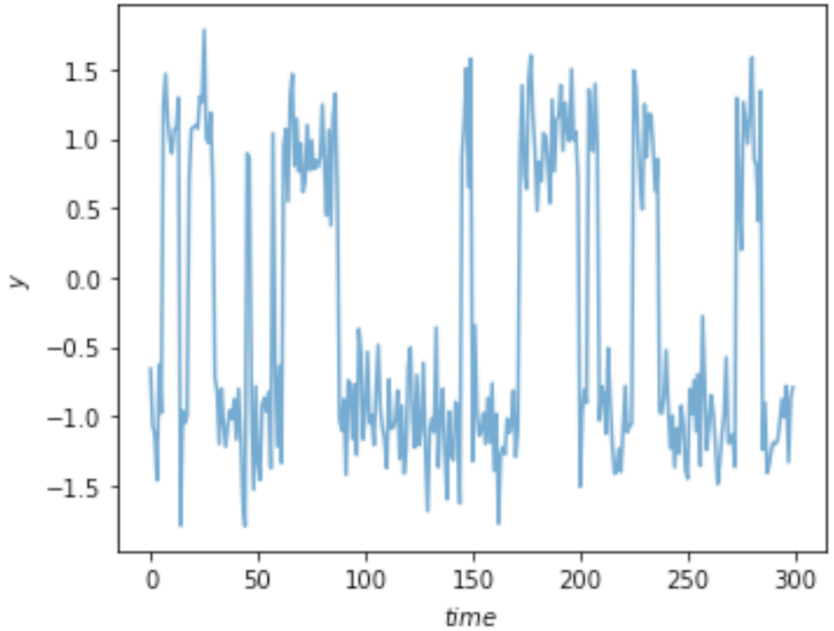
$$\mathbf{r}(t) = (r_i(t))_{i=1,\ldots,D}$$

D-dimensional input data vector that is mean free, i.e.

$$\mathbf{r}(t) = \mathbf{r}(t) - \langle \mathbf{r}(t) \rangle_t$$

Computing the covariance of the data at t = 0 and t=τ which is the lag-time chosen.

$$c_{ij}(\tau) = \langle r_i(t) r_j(t+\tau) \rangle_t$$



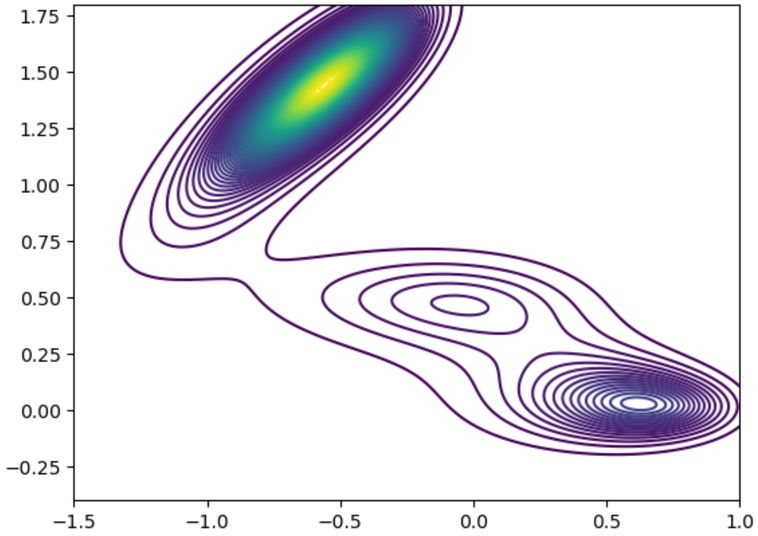
[1]: F. Nüske, B. Keller, G. Pérez-Hernández, **A. S. J. S. Me**y and F. Noé: *J. Chem. Theory Comput.* **10**, 1739-1752 (2014).

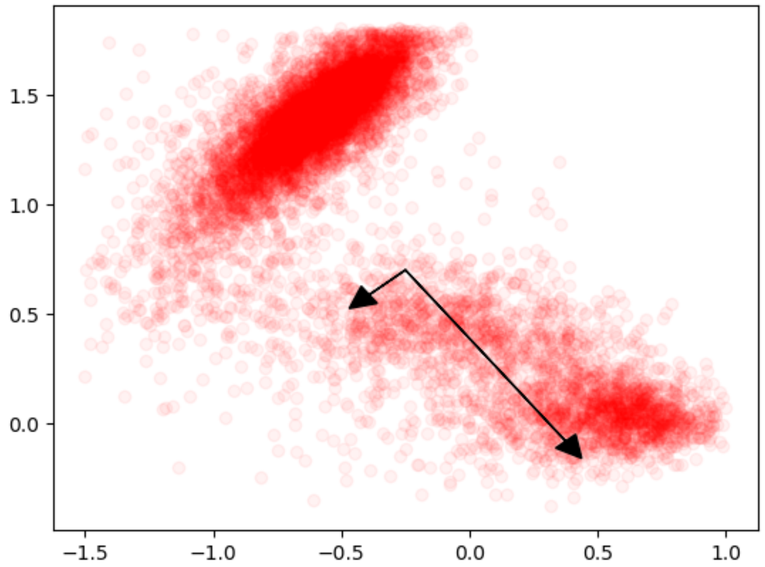
[2]: G. Perez-Hernandez, et al. *J. Chem. Phys.* **139**, 015102 (2013).

[3]: C. R. Schwantes et al. *J. Chem. Theory Comput.* **9**, 2000-2009 (2013).

PCA example

- PCA is an **orthogonal linear transformation** that **maximises the variance** across the first component
- A linear regression fit **minimises the error** with regard to all data points.
- PCA can be used as a tool for **dimensionality reduction**



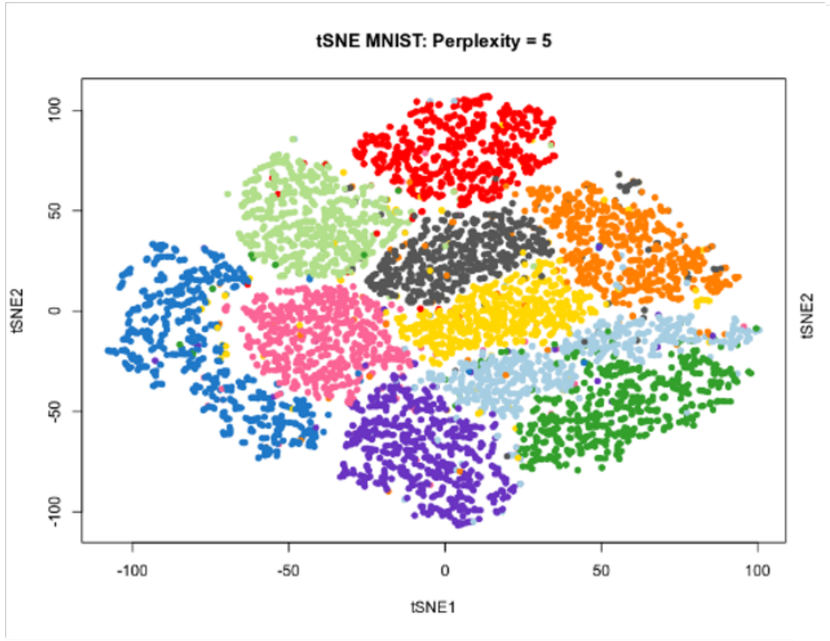


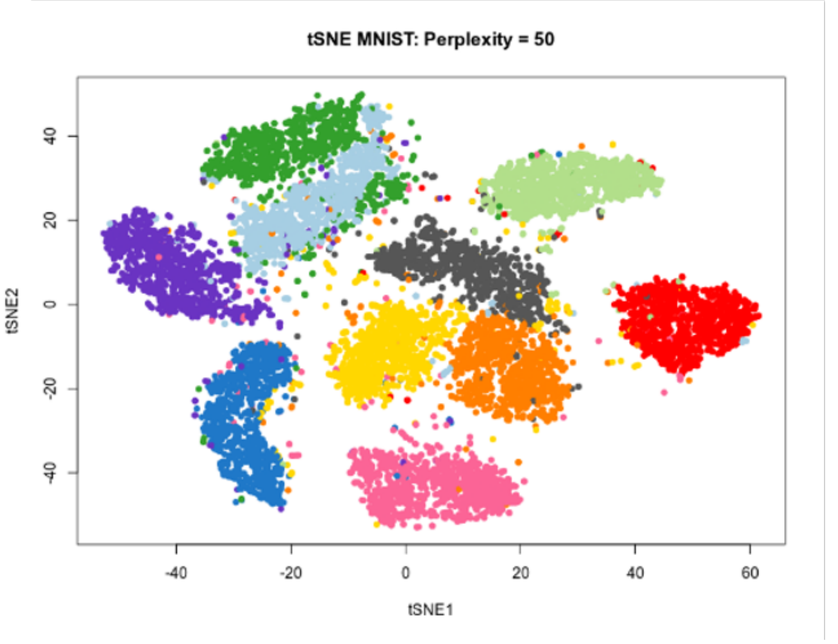
Finding the dominant reaction coordinate of a potential energy surface of a reaction!

T-distributed Stochastic Neighbour Embedding (t-SNE)

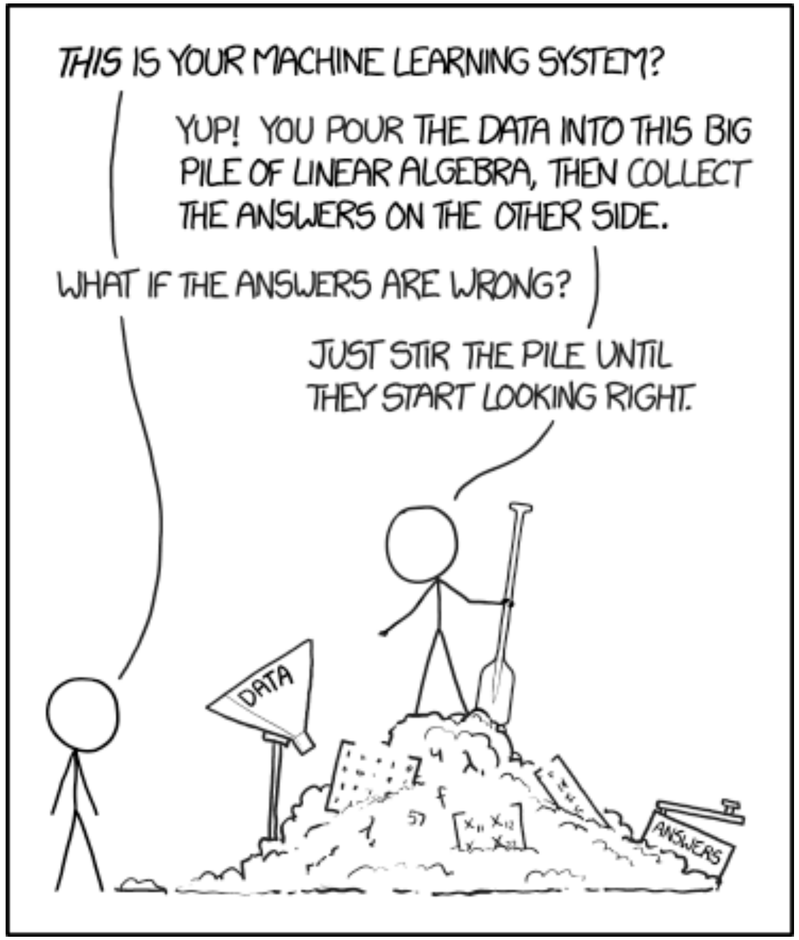
- Useful for visualisation, project high-dimensional data in 2 or 3 dimensions.
- Controlled by one main parameters: “perplexity”
- Relative distance between points not quantitatively meaningful

MNIST: database of written digits





Valuable lessons on machine learning



https://xkcd.com/1838/

Something you liked 😎👍

Something you think could be improved 😬👎